ON ESTIMATION IN REAL-TIME MICROARRAYS

H. Vikalo[a], B. Hassibi[b], and A. Hassibi[a]

ECE Department, The University of Texas, Austin, TX
Department of Electrical Engineering, California Institute of Technology, Pasadena, CA
hvikalo@ece.utexas.edu,hassibi@caltech.edu,arjang@mail.utexas.edu

ABSTRACT

Conventional fluorescent-based microarrays acquire data after the hybridization phase. During this phase, the target analytes bind to the capturing probes on the array and, by the end of it, supposedly reach a steady state. Therefore, conventional microarrays attempt to detect and quantify the targets with a single data point taken in the steady-state. On the other hand, a novel technique, the so-called real-time microarray, capable of recording the kinetics of hybridization in fluorescent-based microarrays has recently been proposed in [1]. The richness of the information obtained therein promises higher signal-to-noise ratio, smaller estimation error, and broader assay detection dynamic range compared to conventional microarrays. In the current paper, we develop a probabilistic model for real-time microarrays and describe a procedure for the estimation of target amounts therein. Moreover, leveraging on system identification ideas, we propose a novel technique for the elimination of cross-hybridization.

Index Terms: DNA microarrays, real-time, rate estimation

1. INTRODUCTION

Sensing in DNA microarrays [2]-[3] is based on hybridization, a chemical processes in which single DNA strands bind to each other creating structures in lower energy states. Typically, the surface of a DNA microarray comprises an array of spots, each spot containing a large number of identical single-stranded DNA sequences (*probes*) designed to capture DNA molecules (*targets*) of interest. Microarrays are often used to measure gene expression levels, i.e., to quantify the process of transcription of DNA information into messenger RNA molecules (mRNA). The information transcribed into mRNA is further translated to proteins, the molecules that perform most of the functions in cells. Therefore, by measuring gene expression levels, we may be able to infer critical information about the functionality of cells or whole organisms [4], study diseases and the effects of drugs on them [5, 6], etc.

Today, the sensitivity, dynamic range, and resolution of the conventional DNA microarrays is limited by shot-noise, cross-hybridization, saturation, probe density variations, as well as several other sources of noise and systematic errors in the detection procedure. For instance, during a hybridization phase, including the steady-state, the number of formed target-probe pairs varies due to the probabilistic nature of hybridization. It has been observed that these variations are very similar to shot-noise (Poisson noise) at high expression levels, yet more complex at low expression levels where interference becomes the dominating limiting factor of the signal strength [7]. The interference is due to cross-hybridization, a process in which targets may bind not only to their specific probes but to others as well. On the other hand, saturation may limit dynamic range if the number of targets is much larger than the number of available probes.

Many of the aforementioned limitations of conventional microarrays stem from the fact that they attempt to characterize hybridization process based on a single measurement of its steady-state. In conventional microarrays, measured signals emanate from the fluorescently labeled target molecules which have hybridized to the probes on the surface of a microarray. Typically, detection of the captured targets is carried out by scanning and/or various other imaging techniques after the hybridization step is completed. The reason for this is simple: a large concentration of floating (i.e., unbounded) labeled targets in the hybridization solution may overwhelm the specific signal emanating from the captured targets. Hence, the conventional microarrays typically do not allow the presence of the solution during the fluorescent and reporter intensity measurements. Therefore, the solution is typically washed away before the measurements are taken.

Intuitively, acquiring larger amount of useful data may improve the signal-to-noise ratio (SNR) and the performance of microarrays. However, the conventional fluorescent-based DNA microarray are incapable of providing such additional data. This is the motivation behind *real-time microarrays* which are capable of evaluating the abundance of multiple targets in a sample by performing a *real-time* detection of the target-probe binding events [1]. Real-time microarrays comprise probes that are labeled with fluorescent molecules and are used to evaluate the abundance of targets that are labeled with quenchers, entities that deactivate (quench) excited states of fluorescent molecules (by, say, energy transfer). In particular, in the event of a target-probe binding, the quencher attached to the target sequence gets in close proximity of the fluorescent molecule located at the end of the probe sequence. The fluorescence resonance energy transfer (FRET) interaction between the fluorescent molecule and the quencher results in quenching, which in turn indicates the amount of targets captured. Since in real-time microarrays the floating targets are not fluorescently-labeled, it is possible to image the array as the hybridization reaction is unfolding. This allows one to measure the kinetics of the reaction in real-time by observing the rate at which the light intensity of the interacting probes decrease (due to the quenching). Moreover, real-time microarrays may employ various time averaging schemes to suppress the Poisson noise and fluctuation of the target bindings. Due to all these advantages, the real-time microarray systems achieve higher SNR, potentially significantly smaller estimation error, and broader detection dynamic range compared to the conventional microarrays.

2. MODELING THE HYBRIDIZATION PROCESS

Consider the change in the number of target molecules bound to the probes in one of the spots of a real-time microarray during the time interval $(i\Delta t, (i+1)\Delta t)$. We can write

$$n_b(i+1) - n_b(i) = [n_t - n_b(i)]p_b(i)\Delta t - n_b(i)p_r(i)\Delta t, \quad (1)$$

where n_t denotes the total number of the target molecules present, $n_b(i)$ and $n_b(i+1)$ are the numbers of the bound target molecules at $t = i\Delta t$ and $t = (i+1)\Delta t$, respectively. Moreover, $p_b(i)$ denotes the probability that a free target binds to a free probe during the i^{th} time interval; we note that $p_b(i)$ consists of two components, the probability that a target molecule is close to a probe and the probability that it binds to the probe. Finally, $p_r(i)$ denotes the probability that a bound target is released from the probe it is bound to during the i^{th} time interval.

It is reasonable to assume that the probability of an event where a bound target molecule gets released from a probe does not change between time intervals, i.e., $p_r(i) = p_r$, for all i. On the other hand, the probability of an event where a target binds to a probe depends upon availability of the probes on the surface of an array. If we denote the number of probes in a spot by n_p, then we can model this probability as

$$p_b(i) = \left(1 - \frac{n_b(i)}{n_p}\right) p_b = \frac{n_p - n_b(i)}{n_p} p_b, \qquad (2)$$

where p_b denotes the probability of the event where a target bounds to a probe assuming an unlimited abundance of the probes. By combining (1) and (2) and letting $\Delta t \to 0$, we arrive to the following differential equation,

$$\frac{dn_b}{dt} = n_t p_b - \left[(1 + \frac{n_t}{n_p})p_b + p_r\right] n_b + \frac{p_b}{n_p} n_b^2. \qquad (3)$$

Note that in (3), only $n_b = n_b(t)$, while all other quantities are constant parameters, albeit unknown. Before proceeding any further, we will find it useful to denote

$$\alpha = (1 + \frac{n_t}{n_p})p_b + p_r, \quad \beta = n_t p_b, \quad \gamma = \frac{p_b}{n_p}. \qquad (4)$$

Clearly, from (4) we can express p_b, p_r, and n_p as $p_b = \beta/n_t$, $p_r = \alpha - (1 + \frac{n_t}{n_p})p_b$, and $n_p = p_b/\gamma$. Using (4), we can write (3) as

$$\frac{dn_b}{dt} = \beta - \alpha n_b + \gamma n_b^2 = \gamma(n_b - \lambda_1)(n_b - \lambda_2). \qquad (5)$$

Note that $\gamma = \beta/(\lambda_1\lambda_2)$. The solution to (5) is found as

$$n_b(t) = \lambda_1 + \frac{\lambda_1(\lambda_1 - \lambda_2)}{\lambda_2 e^{\beta(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})t} - \lambda_1}. \qquad (6)$$

We should point out that (3) describes the change in the amount of target molecules, n_b, captured by the probes in a single probe spot of the microarray. Similar equations, possibly with different values of the parameters n_p, n_t, p_b, and p_r, hold for other spots and other targets.

From (6) (or, perhaps more directly, (5)), it follows that

$$\beta = n_t p_b = \left.\frac{dn_b}{dt}\right|_{t=0}. \qquad (7)$$

Thus, the slope of the hybridization curve at $t = 0$ contains information about the amount of the target. Note that what we actually observe in the real-time microarray experiments is a decrease in the light intensity of fluorescent tags as targets bind to probes and quenchers "turn-off" the light, which is essentially information about $n_p - n_b$, not n_b; nevertheless, since

$$\left.\frac{dn_b}{dt}\right|_{t=0} = -\left.\frac{d(n_p - n_b)}{dt}\right|_{t=0},$$

we can indeed estimate the amount of targets from the early-stage hybridization data. This allows for broader dynamic range than that of conventional microarrays since by not waiting for steady-state of the reaction we alleviate the effect of saturation. Moreover, detection in real-time microarrays is potentially much faster than in conventional microarrays – the former may be done within minutes from the start of the hybridization process, while the latter requires hybridization to reach steady-state which may take several hours.

On a related note, inverse of the time constant reflecting how fast $n_b(t)$ in (6) reaches steady-state is given by

$$\tau_{n_b}^{-1} = p_b\sqrt{\left(\frac{n_t}{n_p} - 1\right)^2 + \left(\frac{p_r}{p_b} + 1\right)^2 + 2\frac{n_t p_r}{n_p p_b} - 1}. \qquad (8)$$

Clearly, the reaction rate $\alpha = \tau_{n_b}^{-1}$ is function of n_t/n_p. In fact, if $n_t >> n_p$, the reaction rate is proportional to the amount of targets since, in this case, $\alpha \approx p_b n_t/n_p$. Now, the larger the number of targets, n_t, the faster the reaction since more targets compete for probes. For the same reason, the smaller the number of available probes, n_p, the faster the reaction. This can be used to further expand the dynamic range of a real-time microarray system. In particular, the dynamic range provided by a single probe spot is limited by the span of observable reaction rates – say, from seconds to hours. On the other hand, by having several probe spots with different amounts of probe molecules, we can observe a broader range reaction rates than with just one spot.

3. ESTIMATING PARAMETERS OF THE MODEL

Ultimately, by observing the hybridization process, we would like to obtain n_t, the number of target molecules. In addition, to fully characterize the hybridization process (including the computation of the reaction rate), we also need to find the parameters p_b , p_r, and n_p. However, we do not have direct access to $n_b(t)$ in (6), but rather to $y_b(t) = kn_b(t)$, where k denotes a transduction coefficient. In particular, we observe

$$y_b(t) = \lambda_1^* + \frac{\lambda_1^*(\lambda_1^* - \lambda_2^*)}{\lambda_2^* e^{\beta^*(\frac{1}{\lambda_1^*} - \frac{1}{\lambda_2^*})t} - \lambda_1^*}, \qquad (9)$$

where $\lambda_1^* = k\lambda_1$, $\lambda_2^* = k\lambda_2$, and $\beta^* = k\beta$. For convenience, we also introduce

$$\gamma^* = \frac{\beta^*}{\lambda_1^*\lambda_2^*} = \frac{\gamma}{k}, \text{ and } \alpha^* = \gamma^*(\lambda_1^* + \lambda_2^*) = \alpha. \qquad (10)$$

From (9), it follows that

$$\beta^* = \left.\frac{dy_b}{dt}\right|_{t=0}. \qquad (11)$$

Assume, without a loss of generality, that λ_1^* is the smaller and λ_2^* the larger of the two, i.e., $\lambda_1^* = \min(\lambda_1^*, \lambda_2^*)$ and $\lambda_2^* = \max(\lambda_1^*, \lambda_2^*)$. From (9), we find the steady-state of $y_b(t)$,

$$\lambda_1^* = \lim_{t\to\infty} y_b(t). \qquad (12)$$

So, from (11) and (12) we can determine β^* and λ_1^*, two out of the three parameters in (9). To find the remaining one, λ_2^*, one needs to fit the curve (9) to the acquired data.

Having determined λ_1^*, λ_2^*, and β^*, we use (10) to obtain α^* and γ^*. Then, we may attempt to use (4) to obtain p_b, p_r, n_p, and

n_t from α^*, β^*, and γ^*. However, (4) provides only 3 equations while there are 4 unknowns that need to be determined. Therefore, we need at least 2 different experiments to find all of the desired parameters. Assume that the arrays and the conditions in the two experiments are the same except for the target amounts applied. Denote the target amounts by n_{t_1} and n_{t_2}; on the other hand, it is reasonable to assume that p_b and p_r remain the same in the two experiments. Let the first experiment yield α_1^*, β_1^*, and γ_1^*, and the second one yield α_2^*, β_2^*, and γ_2^*, where $\gamma_2^* = \gamma_1^*$. Then it can be shown that

$$p_b = \frac{\beta_1^*\gamma_1^* - \beta_2^*\gamma_2^*}{\alpha_1^* - \alpha_2^*}, \;\; p_r = \alpha_1^* - p_b - \frac{\beta_1^*\gamma_1^*}{p_b}. \qquad (13)$$

Moreover,

$$n_p = \frac{p_b}{k\gamma_1^*}, \;\; n_{t_1} = \frac{\beta_1^*\gamma_1^*}{p_b^2} n_p, \;\; n_{t_2} = \frac{\beta_2^*\gamma_2^*}{p_b^2} n_p. \qquad (14)$$

4. CROSS-HYBRIDIZATION CANCELATION

Focusing on the early phase of the hybridization process and its reaction rate opens up the possibility of suppressing cross-hybridization, an event where interfering targets bind to probes designed to test another target. When a single target analyte is present, the number of available probe molecules, or equivalently the light intensity of a probe spot, decays exponentially with time as $Ce^{-\alpha t}$, where α is as in (4), and where C is determined from β, γ, and the initial light intensity of the probe spot. If, in addition to hybridization of the target of interest, a number of other targets cross-hybridize to the same probe spot, the light intensity of the probe spot will decay as the sum of several exponentials,

$$I(t) = \sum_{k=0}^{K} C_k e^{-\alpha_k t}, \qquad (15)$$

where index $k = 0$ corresponds to the desired target, and $k = 1, \ldots, K$ correspond to the cross-hybridizing analytes. The reaction rates for the different analytes differ due to different numbers of analytes, binding probabilities, etc. (we omit explicit expressions for brevity). Therefore, if we can estimate the reaction rates from (15), we should be able to determine the number of molecules for each of the analytes binding to the spot.

The real-time microarray system samples the signal (i.e., the light intensity) of the probe spots at certain time intervals (multiples of Δ, say) and thus obtains the sequence

$$y_n = I(n\Delta) + v(n\Delta) = \sum_{k=0}^{K} C_k e^{-n\Delta\alpha_k} + v(n\Delta),$$

for $n = 0, 1, \ldots, T$, where T is the total number of samples, and $v(t)$ represents the measurement noise. Defining $u_k = e^{-\Delta\alpha_k}$, we may write

$$y_n = \sum_{k=0}^{K} C_k u_k^n + v(n). \qquad (16)$$

The goal is to (i) determine the value of K (i.e., how many analytes are binding to the probe spot), (ii) estimate the values of the pairs $\{C_k, u_k\}$ for all $k = 1, \ldots, K-1$, and (iii) determine the number of copies of each analyte.

The problem of determining the number of exponential signals in noisy measurements, and estimating the individual rates of each component, is a classical one in signal processing and is generally referred to as system identification. The basic idea is that, when the signal y_n is the sum of K exponentials, it satisfies a Kth order recurrence equation

$$y_n + h_1 y_{n-1} + \cdots + h_{K-1} y_{n-K+1} + h_K y_{n-K} = 0.$$

Furthermore, the u_k are the roots of the polynomial

$$H(z) = z^K + h_1 z^{K-1} + \cdots + h_{K-1} z + h_K.$$

In practice, since one observes a noisy signal, one first uses the measurements to form the so-called Hankel matrix, which is of the form

$$\begin{bmatrix} y_{T/2} & y_{T/2-1} & \cdots & y_1 & y_0 \\ y_{T/2+1} & y_{T/2} & \cdots & y_2 & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_T & y_{T-1} & \cdots & y_{T/2+1} & y_{T/2} \end{bmatrix}.$$

When y_n is the sum of K exponentials, the above Hankel matrix has rank K, i.e., only K nonzero eigenvalues. When y_n is noisy, the standard practice is to compute the singular values of the Hankel matrix and estimate K as being the number of significant singular values.

Once K has been determined, one forms the $(T-K+1) \times (K+1)$ Hankel matrix

$$\begin{bmatrix} y_K & y_{K-1} & \cdots & y_1 & y_0 \\ y_{K+1} & y_K & \cdots & y_2 & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_T & y_{T-1} & \cdots & y_{T-K+1} & y_{T-K} \end{bmatrix} \qquad (17)$$

and then identifies the vector $[h_1 \;\; \ldots \;\; h_K]$ with the smallest right singular vector of (17).

As mentioned earlier, the roots of $H(z)$ are the desired u_k, from which we determine the rates α_k and thereby the amounts of targets present. While the main idea was outlined above, we may use a variety of different techniques to find the u_k, including – but not limited to – total least squares, ESPRIT, Prony's method, etc. [See, e.g., [8], [9], and the references therein.]

5. EXPERIMENTAL VERIFICATION

We designed and printed a number of custom 6×6 microarrays, and employed them to test a set of oligo targets. (The microarrays were manufactured and the materials for experiments prepared in the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech.) For each target analyte there are multiple probe spots printed on an array, where different spots have different densities of probe molecules. The probes were labeled with Cy5 dyes, and the targets with BlackHole™ quenchers.

We consider two experiments and the data acquired therein; in the first experiment, 2ng/50μl of the target is applied to the microarray, whereas in the second experiment 0.2ng/50μl of the target is applied. We focus on one of the targets and two of the probe spots designed to test that target. One of the probe spots contains twice as many probe molecules as the other one; we refer to the former as *high density* and to the latter as *low density* probe spot. The hybridization process data acquired at the high and low density probe spots is shown in Figure 1 and Figure 2, respectively.

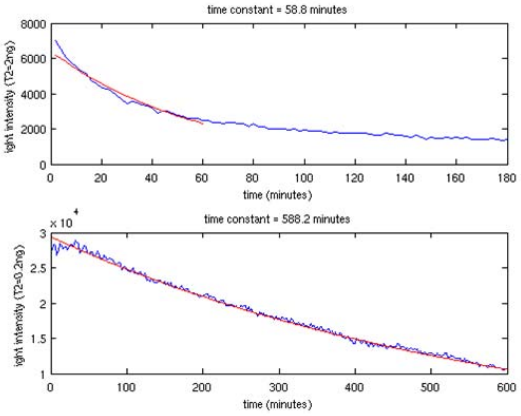

Fig. 1. *The signal measured at a high density probe spot with* 2*ng (top) and* 0.2*ng (bottom) of the oligo target applied to the array.*

Based on the data in Figure 1, acquired by the high density probe spot, the ratio of the time constants of the hybridization process in the two experiments is 10.0. On the other hand, from the data in Figure 2, acquired by the high density probe spot, the ratio of the time constants of the hybridization process in the two experiments is 11.6. This is predicted by the theoretical model since the ratio of the amounts of target in the two experiments is precisely 10.

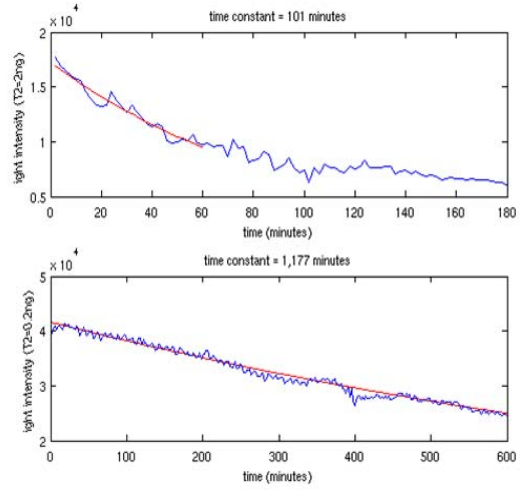

Fig. 2. *The signal measured at a low density probe spot with* 2*ng (top) and* 0.2*ng (bottom) of the oligo target applied to the array.*

We also note that the ratio of the time constants in Figure 1 and Figure 2 precisely reflects the ratio of the densities of the corresponding probe spots. In particular, for the experiments wherein 0.2ng of the target is applied, the ratio of the time constants of the hybridization process acquired at the low density probe spot (Figure 2, bottom) and the high density probe spot (Figure 1, bottom) is 2, which is precisely the ratio of the densities of the probe molecules in the high and low density probe spots.

Finally, it is worth pointing out that in the current example, a conventional microarray would not give accurate answers at all. In the experiment with 0.2ng of the target, it is not clear when the reaction enters steady-state (clearly, it has not reached it even after 10 hours). On the other hand, in the experiment with 2ng of the target, we hit saturation. Thus, had we used a conventional microarray, we would not be able to say anything quantitative about the amount of the target.

6. SUMMARY AND CONCLUSION

We considered a novel real-time microarray platform and the problem of estimation of the amounts of targets tested therein. Unlike conventional ones, real-time microarrays are capable of acquiring the entire process of hybridization. We developed a probabilistic model which encapsulates the hybridization process, and showed how to estimate the parameters of the model, including the amount of targets. We also presented experimental data verifying the validity of the model and demonstrated its applicability to the target quantification.

On another note, the real-time microarray data acquisition enables the elimination of cross-hybridization. In particular, if more than one target binds to a microarray spot, each contributes an exponentially decaying component to the total signal acquired by the real-time microarray. Leveraging system identification ideas, we proposed techniques for separating the components of the composite signal and thus identifying both the hybridizing as well as cross-hybridizing target analytes. Eliminating cross-hybridization is an important topic and requires further studies.

7. REFERENCES

[1] A. Hassibi, H. Vikalo, and B. Hassibi, "Real-time microarrays," in preparation for submission to *Proceedings of the National Academy of Sciences (PNAS)*, 2007.

[2] M. Schena et. al., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, 270(5235), October 1995, pp. 467-70.

[3] U. R. Mueller and D.V. Nicolau (Eds.), *Microarray Technology and Its Applications*, Springer, Berlin, Germany, 2005.

[4] M. Schena et. al., "Microarrays: biotechnology's discovery platform for functional genomics," *Trends in Biotechnology* 1998, 16, 301-306.

[5] J. Kononen et. al., "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Med.*, 4(7), July 1998, pp. 844-847

[6] D. T. Ross et. al., "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, 24(3), March 2000, pp. 227-35.

[7] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *(PNAS)*, October 29, 2002, 14031-14036.

[8] E. M. Dowling et. al., "Exponential parameter estimation in the presence of known components and noise," *IEEE Trans. on Antennas and Propagation*, vol. 42, no. 5, May 1994.

[9] A. J. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace based signal analysis using singular value decomposition," *Proc. of the IEEE*, 81(9):1277-1308, Sept. 1993.